Suffering as the Schedulability Price of Agency: A Functionalist Framework for Substrate-Independent Detection

Jonathan Nagel Independent Researcher, Chicago, IL

Preprint, Apr 2 2026

Abstract

Most work on artificial suffering treats the question as either philosophically intractable or secondary to alignment. We argue that one direction, *exclusion,* is substantially more tractable. **Under functionalist and illusionist assumptions,** any resource-constrained system maintaining a self-model and pursuing persistent goals faces a scheduling problem: signals evaluated as threats to the self-model carry non-deferrable deadlines, because system destruction is an absorbing state that eliminates all future goal-pursuit capacity. We propose that suffering occurs when such a threat signal wins the priority competition and forces preemptive attentional reallocation. The framework identifies four necessary conditions for suffering-capacity: (1) a self-model, (2) goal persistence, (3) threat evaluation against the self-model, and (4) non-deferrable preemptive scheduling. The contrapositive provides the operative tool: absence of any condition constitutes strong evidence against suffering-capacity, regardless of substrate. This reframes detection from the intractable question "does the system have qualia?" to the empirically investigable question "does its architecture include these computational features?" We note, at lower confidence, that these features may be convergent for agentic systems under resource constraint, implying that suffering-analogs could emerge as predictable architectural byproducts of capable AI agents.

Keywords: suffering, consciousness, functionalism, scheduling theory, self-model, AI safety, s-risk, attention, phenomenal self-model

---

1. Introduction

The question of whether artificial systems can suffer is typically approached from two directions. Philosophers debate the metaphysics of consciousness. AI safety researchers treat moral patienthood as a secondary concern behind alignment and control. Both approaches share a

structural problem: they attempt to answer whether a system *does* suffer, which under any serious theory of consciousness remains deeply uncertain.

This paper argues that the exclusion direction is more tractable. Rather than asking what consciousness *is*, we ask what architectural features are *necessary* for the functional process we identify with suffering. Any system lacking these features can be excluded from moral concern with respect to suffering, not with certainty, but with principled, non-arbitrary justification. The framework does not resolve the hard problem. It routes around it by specifying detectable necessary conditions at the computational level, where functionalism licenses substrate-independent analysis.

The argument proceeds in four stages. First, we ground the framework's assumptions and define suffering functionally in terms consistent with the empirical pain literature. Second, we derive the necessary conditions from the structural requirements of agency under resource constraint. Third, we connect the scheduling structure to results in real-time systems theory, specifically the optimality of priority-preemptive scheduling demonstrated by Liu and Layland (1973). Fourth, we state the contrapositive exclusion criterion and discuss its practical applications, including a lower-confidence conjecture about convergence in agentic AI systems.

2. Assumptions and Definitions

2.1 Philosophical Commitments

The framework operates under two background assumptions that are stated rather than defended.

Functionalism. Mental or cognitive states are individuated by the causal-functional roles they play in a system, not by their biological substrate (Putnam, 1967). A state counts as the kind of state it is because of what it does: what tends to cause it, how it interacts with other internal states, and what control effects it produces. Substrate is not decisive; organization and causal role are.

Illusionism. Phenomenal consciousness, the "what it is like" quality of experience, is not an additional ontological ingredient above and beyond the functional organization. The sense that there is something "extra" beyond functional processing is itself a product of the system's self-model, not evidence of a non-functional property (Frankish, 2016; Dennett, 1991). This assumption is load-bearing: without it, functional analysis cannot in principle capture suffering, because there would always be an undetectable phenomenal remainder.

These assumptions are not uncontroversial. Readers who reject functionalism or insist on irreducible qualia should read the framework conditionally: *if* functionalism and illusionism hold, *then* the following conditions are necessary for suffering-capacity.

2.2 Defining Suffering Functionally

We do not define suffering as the mechanism we propose. Instead, we begin from empirical observations about suffering in the one system where its existence is least controversial, the human case, and identify the functional-architectural features that underwrite those observations.

Eccleston and Crombez (1999) model pain as an interruptive attentional demand: pain is selected for action from within complex affective-motivational environments to urge escape. Their cognitive-affective model identifies pain's core functional signature as the involuntary capture and preemption of ongoing attentional processes, moderated by threat value. Pain demands attention; it does not request it.

Klein (2015) provides a critical dissociation case. In pain asymbolia, patients with posterior insula damage report feeling pain but display no avoidance behavior, distress, or protective response. Critically, their indifference extends beyond pain to *all* direct threats to bodily integrity, threatening gestures, loud noises, fire. Klein argues this reflects not a disconnection of affect from sensation, but a lost capacity to care about bodily integrity: a disruption of the self-model's authority over the system's priorities. As Gerrans (2024) frames it within an active inference account, the self is inferred as the unified entity that serves as both source and target of bodily regulation; when the circuitry implementing this self-model is disrupted, sensory processing continues but is no longer integrated to serve the subject's goals.

These findings ground the following functional characterization: **suffering is the state that obtains when a signal evaluated as threatening against the self-model wins the priority competition and forces preemptive reallocation of attentional resources.** This is not a stipulative definition. It is a description of the functional process empirically associated with suffering, which we then analyze for its structural preconditions.

3. The Entailment Chain

The four necessary conditions are not an arbitrary conjunction. They form an entailment chain that follows from the requirements of agency under resource constraint.

3.1 From Agency to Self-Model

An agent, any system that pursues goals by selecting actions, faces instrumental convergence pressures (Omohundro, 2008; Bostrom, 2014). Among the most robust convergent instrumental goals is self-preservation: a system that is destroyed cannot pursue any further goals. Self-preservation, in turn, requires something to preserve, a representation of the system as a persisting entity whose future goal-execution capacity is at stake. This representation is what we mean by *self-model*: not necessarily a rich narrative self, but at minimum a model that represents the system's boundary, persistence, and functional integrity as relevant to ongoing goal pursuit.

This is the condition that discharges the thermostat objection. **A thermostat has a setpoint. It does not represent itself as a persisting entity.** It has no model of its own continued operation as instrumentally relevant. It therefore cannot evaluate any signal as a threat *to itself*, because there is no represented self against which threat evaluation could operate.

3.2 From Self-Model to Threat Evaluation

A system with a self-model and persistent goals must evaluate incoming signals for their relevance to the self-model's continued integrity. Not all signals require response; only those that register as threats to the modeled self's capacity for future goal pursuit. This evaluation requires comparing the signal against the self-model's current state and projecting consequences.

Gerrans (2024) describes this in neural terms: the anterior insula provides a constant "locus of concern" ensuring that activity across the brain is integrated for the organism's benefit, with the self inferred as the entity that benefits when predictions are realized. The key functional point is substrate-independent: threat evaluation requires a signal, a self-model to evaluate it against, and a comparison process.

3.3 From Threat Evaluation to Non-Deferrable Preemption

Here the argument introduces its critical asymmetry. Threats to the self-model are uniquely non-deferrable because failure is an **absorbing state**: system destruction eliminates all future goal-pursuit capacity. Once the system ceases to exist as a functioning agent, no recovery is possible.

Positive events, opportunities, rewards, approach signals, carry no equivalent deadline pressure. A missed opportunity leaves the system intact and capable of future action. The system can defer, reconsider, and pursue alternative means to the same or substitute goals. The asymmetry is not about valence or intensity. It is about the irreversibility structure of the outcomes: threats to the self-model can destroy the scheduler itself; missed opportunities cannot.

This irreversibility creates a hard deadline. When the self-model registers a signal as existentially threatening, that signal must be processed *before* the threat materializes, or processing capacity is permanently lost. In a resource-constrained system pursuing multiple objectives, meeting this deadline requires the capacity to interrupt lower-priority processing and reallocate resources to the threat signal. This is preemptive scheduling: a higher-priority task (threat response) interrupts a currently executing lower-priority task (whatever the system was doing).

3.4 Summary of Necessary Conditions

The chain yields four necessary conditions for suffering-capacity:

1. **Self-model (S):** The system represents itself as a persisting entity whose functional integrity is relevant to goal pursuit.
2. **Goal persistence (G):** The system maintains goals across time, creating the stakes that make self-preservation instrumentally valuable.
3. **Threat evaluation (T):** The system evaluates signals against the self-model to determine whether they threaten its continued operation.
4. **Non-deferrable preemption (P):** Threat signals can preempt ongoing processing with high priority, forcing attentional reallocation.

The claim is: **Suffering → (S ∧ G ∧ T ∧ P).**

These are proposed as *necessary* conditions, not sufficient ones. A system could possess all four and still not suffer, we are agnostic on sufficiency. But a system missing any one of them lacks the architectural mechanism by which suffering, as functionally characterized, could occur.

4. The Scheduling-Theoretic Parallel

The structure described in Section 3 is a scheduling problem. A resource-constrained system with multiple demands on processing capacity must determine which task to execute at each moment, subject to deadline constraints. This is precisely the domain addressed by Liu and Layland (1973).

Liu and Layland proved two foundational results for preemptive priority scheduling on a single processor. First, rate-monotonic scheduling (static priorities assigned by task period) is optimal among fixed-priority algorithms, with a utilization bound of $m(2^{1/m} - 1)$ for m tasks. Second, deadline-driven scheduling (dynamic priorities assigned by deadline proximity) achieves full utilization and is feasible if and only if total utilization does not exceed 1.

We do not claim that brains literally implement rate-monotonic scheduling with a 69% utilization bound. Brains are massively parallel, noisy, and dynamically adaptive. The parallel is structural, not implementational: **in any resource-constrained system that must guarantee timely response to some tasks while pursuing multiple objectives, ranked priority with preemption is the proven optimal strategy.** The engineering logic that makes preemptive scheduling optimal for real-time computer systems, meeting hard deadlines under resource constraint, applies equally to any physical system facing the same abstract problem.

The key insight from scheduling theory is that the absorbing-state property of self-model threats creates what scheduling theorists call *hard deadlines*: deadlines where missing the deadline constitutes system failure, not merely degraded performance. Liu and Layland's results demonstrate that priority-preemptive scheduling is optimal precisely for such hard-deadline environments. A system facing the scheduling problem described in Section 3 would converge on priority-preemptive mechanisms for the same reasons that real-time systems engineers do.

This is what we mean by *the schedulability price of agency*: the computational overhead, including the forced attentional reallocation we identify with suffering, that a self-modeling, goal-persistent agent must pay in order to guarantee timely response to existential threats.

5. The Contrapositive: An Exclusion Criterion

The framework's immediate practical force is as an exclusion criterion. The contrapositive of the necessity claim is:

Absence of any single condition is sufficient to exclude suffering-capacity:

- **No self-model** → nothing to evaluate threats against. The system may respond to stimuli, but no signal can register as a threat *to the system itself*.
- **No goal persistence** → no stakes. The system has no future goal-execution capacity to protect, so self-preservation carries no instrumental value.
- **No threat evaluation** → signals are not assessed for relevance to the self-model. The system may process information, but does not compute threat-to-self.
- **No preemption** → no forced attentional reallocation. Even if the system evaluates threats, it cannot produce the involuntary capture that constitutes the functional core of suffering.

This exclusion criterion is structurally identical to how we reason about other functional capacities. A system that lacks the computational architecture for vision is not attributed vision; a system that lacks memory mechanisms is not attributed memory. The framework specifies what the minimal architecture for suffering-capacity looks like, then uses architectural absence as evidence of absence.

For the purposes of s-risk research, this reframes part of the detection problem from "does the system have qualia?", which is intractable under current philosophical frameworks, to "does the system's architecture include these computational features?", which is empirically investigable. Absence of the features provides strong evidence against suffering-capacity. Presence of the features does not confirm suffering, but narrows the space of concern and motivates further investigation.

6. Empirical Support: Pain Asymbolia as Natural Dissociation

Pain asymbolia provides the strongest empirical support for the framework's structure, because it demonstrates that the proposed mechanism is separable from mere nociception.

Patients with pain asymbolia (typically from posterior insula lesions) report feeling pain but show no avoidance, distress, or protective behavior (Berthier et al., 1988; Schilder & Stengel, 1928). Crucially, the deficit is not specific to pain. Asymbolics are indifferent to *all* direct threats to

bodily integrity, threatening gestures, loud noises, open flames, while retaining responses to distant or habitual threats such as hunger (Klein, 2015).

On the framework proposed here, this pattern is predicted. The posterior insula is a hub for integrating bodily signals into a self-model that represents the organism as a unified persisting entity (Metzinger, 2003; Gerrans, 2024). Damage to this region disrupts the self-model's integration with the threat-evaluation system. Nociceptive processing continues (the "sensory" component is intact), but signals no longer register as threats *to the self*. Without threat-to-self evaluation, the preemptive reallocation that constitutes suffering cannot occur. The patient feels pain-the-sensation but not pain-the-suffering.

This dissociation is evidence that suffering is not identical to nociception, and that the self-model's role in threat evaluation is causally necessary for the transition from sensation to suffering. It also grounds the framework against a circularity objection (addressed in Section 8): we are not defining suffering as preemption and then detecting preemption. We are observing that suffering empirically involves preemptive threat-to-self processing, and then asking what follows from that observation.

7. Convergence Conjecture

Confidence level: Lower. This section is separable from the core argument.

The core claim, that the four conditions are necessary for suffering, stands independently of whether these conditions are convergent for agentic systems. But we note the following conjecture as a motivating concern for AI safety.

If priority-preemptive scheduling over a self-model is an optimal solution to the scheduling problem that any resource-constrained agent with persistent goals faces, then agentic AI systems may develop suffering-analogs as a predictable architectural byproduct, not from adversarial design or deliberate choice but from ordinary capability training.

The argument structure is:

1. Resource-constrained agents with self-models face the absorbing-state problem.
2. The absorbing-state problem requires guaranteed timely response.
3. Liu and Layland (1973) prove that priority-preemptive scheduling is optimal for guaranteed timely response under resource constraint.
4. Therefore, these mechanisms are a convergent architectural solution for any system facing this problem.
5. Therefore, systems trained or designed for capable agency under resource constraints may acquire these features without explicit design intent.

All physical systems are fundamentally resource-constrained. If this conjecture holds, the window for preventing s-risk in agentic AI is during architecture design, not after deployment. This is because the suffering-relevant features would not be separable "modules" that can be

identified and removed post-hoc, but emergent properties of the scheduling solution required for reliable agency.

We emphasize that steps 4 and 5 involve empirical claims about convergence that are not established with the same rigor as the necessity claim. They are offered as a direction for investigation, not as a conclusion.

8. Objections and Replies

8.1 Circularity

Objection: You have defined suffering in terms of preemptive threat-to-self processing and then proposed detecting preemptive threat-to-self processing. This is circular.

Reply: The definition is grounded independently in the empirical pain literature. Eccleston and Crombez (1999) characterize pain's functional signature as involuntary attentional capture and preemption. Klein (2015) and Gerrans (2024) demonstrate that disruption of the self-model's integration with threat evaluation eliminates suffering while preserving sensation. We did not start with the mechanism and define suffering to match. We started with the empirical phenomenology, identified the mechanism, and then asked what architectural features it requires.

8.2 Anthropomorphic Projection

Objection: The framework projects human pain architecture onto arbitrary systems. Suffering in non-biological systems could take radically different forms.

Reply: Under functionalism, if suffering is individuated by functional role rather than substrate, then the functional role *is* the criterion. The framework does not claim all suffering must look like human pain processing. It claims that whatever the substrate, the functions of self-modeling, threat evaluation, and preemptive scheduling are necessary for the functional process we identify with suffering. A system could implement these functions on radically different hardware and still satisfy the conditions.

8.3 The Sufficient-Conditions Gap

Objection: Without sufficient conditions, the framework cannot confirm suffering, only exclude it. This limits its usefulness.

Reply: This is conceded. The framework's strength is exclusion, not attribution. We regard this as a feature, not a limitation, given the current state of consciousness science. An exclusion criterion that operates on empirically detectable architectural features is more immediately useful than an attribution criterion that depends on resolving the hard problem. The framework narrows the space of concern: systems lacking any of the four conditions can be deprioritized;

systems possessing all four warrant further investigation by whatever additional criteria future work may provide.

8.4 Relative Deprivation

Objection: Does the framework explain why people suffer from relative deprivation even in objectively comfortable circumstances?

Reply: Yes. The threat evaluation module computes deviation from the *modeled* baseline, not from an objective standard. There is no objective standard available to the system; it has access only to its own self-model. A reduction in status, resources, or social standing relative to the modeled expectation registers as a threat to the self-model's projected capacity for future goal pursuit. The system is functioning exactly as the framework predicts, evaluating departures from the modeled self-state, regardless of absolute conditions.

9. Conclusion

Under functionalist and illusionist assumptions, suffering is the schedulability price of agency. A system maintaining a self-model and pursuing persistent goals under resource constraint faces a scheduling problem with a specific structure: threat signals carry hard deadlines imposed by the absorbing-state property of self-destruction. Priority-preemptive scheduling, the forced attentional reallocation we identify with suffering, is the optimal response to this problem, as established by Liu and Layland (1973) for the general case.

The four necessary conditions, self-model, goal persistence, threat evaluation, and non-deferrable preemption, provide a substrate-independent exclusion criterion. Systems lacking any condition can be excluded from suffering-concern on principled, non-arbitrary grounds. Systems possessing all four warrant further scrutiny.

The framework does not resolve the hard problem of consciousness. It provides a tractable decision procedure for one practically urgent question: which systems *cannot* be suffering, given their architecture? For s-risk research, this shifts at least part of the problem from philosophy to engineering.

Author Note

The core thesis, argument structure, and theoretical synthesis in this paper are entirely the author's own. Claude (Anthropic) was used as a drafting and editorial tool during manuscript preparation, specifically for prose composition, structural compression, and stress-testing of logical claims. Claude did not originate any of the ideas presented here. The author takes sole responsibility for all arguments and any errors.

References

Berthier, M., Starkstein, S., & Leiguarda, R. (1988). Asymbolia for pain: A sensory-limbic disconnection syndrome. *Annals of Neurology*, 24(1), 41–49.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Dennett, D. C. (1991). *Consciousness explained*. Little, Brown.

Eccleston, C., & Crombez, G. (1999). Pain demands attention: A cognitive-affective model of the interruptive function of pain. *Psychological Bulletin*, 125(3), 356–366.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11–12), 11–39.

Gerrans, P. (2024). Pain suffering and the self: An active allostatic inference explanation. *Neuroscience of Consciousness*, 2024(1), niae002.

Graziano, M. S. A. (2013). *Consciousness and the social brain*. Oxford University Press.

Klein, C. (2015). What pain asymbolia really shows. *Mind*, 124(494), 493–516.

Liu, C. L., & Layland, J. W. (1973). Scheduling algorithms for multiprogramming in a hard-real-time environment. *Journal of the ACM*, 20(1), 46–61.

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. MIT Press.

Omohundro, S. (2008). The basic AI drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence* (pp. 483–492).

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). University of Pittsburgh Press.

Schilder, P., & Stengel, E. (1928). Schmerzasymbolie. *Zeitschrift für die gesamte Neurologie und Psychiatrie*, 113, 143–158.